Journal of the American College of Cardiology Vol. 55, No. 5, 2010

Published by Elsevier Inc.
ISSN 0735-1097/10/$36.00
doi:10.1016/j.jacc.2009.06.066

Randomized Trials, Statistics, and Clinical Inference

Gregg W. Stone, MD,* Stuart J. Pocock, PHD†

New York, New York; and London, United Kingdom

The completion and proper assessment of prospective, randomized controlled trials is essential for best medical practice. However, even though randomized trials are generally considered the pinnacle of evidence-based medicine, they are not infrequently poorly designed, implemented with inadequate quality control, and/or are subject to inappropriate interpretation or generalization, resulting in suboptimal clinical care and/or future investigative directions. The present report describes the most common and egregious misrepresentations from randomized trials, many of which may be attributed to the fallacies that arise from underpowered studies, resulting in overly optimistic or unwarranted conclusions. Caution is necessary when assessing composite outcomes, secondary end points, subgroup analyses, and the results of meta-analysis and meta-regression. Sponsors and investigators must accept responsibility for optimizing the design and execution of clinical trials, and practitioners, guidelines committees, editors, and regulators must critically interpret the data and literature arising from such studies. It is hoped that the principles embodied in the present commentary will spur improved design of future randomized trials and thoughtful critical appraisal by health care providers. (J Am Coll Cardiol 2010;55:428–31)

"A p value is no substitute for a brain."
—Anonymous

The practice of medicine comprises endless choices by physicians on behalf of our patients. Such decisions are based on knowledge gained from randomized controlled trials (RCTs), nonrandomized studies, personal and shared experiences, and common sense. By avoiding biases in patient selection between experimental and control groups, the prospective RCT is considered the highest level of scientific evidence (1). RCTs also typically have the greatest regulatory oversight and most robust study processes (e.g., on-site data monitoring, independent event adjudication and data safety monitoring, and blinded core laboratories). However, RCTs are expensive, labor intensive, and time-consuming, and the results apply only to the patients enrolled and to the specific drugs, devices, and procedures tested. Relatively few questions are addressed by RCTs, and the standard of care evolves, rendering some RCTs in part obsolete. Many decisions are, therefore, based on observational evidence and personal experience (anecdotal medicine). Unfortunately, because statistical methods cannot completely adjust for unmeasured confounders, such nonrandomized studies may yield erroneous conclusions (Fig. 1) (2,3), and thus should not generally be used for comparative effectiveness analysis. However, observational registries in large, unselected populations are useful to document practice patterns and quantify the incidence of low frequency events.

Despite the advantages of RCTs, their design, implementation, analysis, reporting, and subsequent clinical inferences can sometimes be seriously flawed. Accordingly, physicians should interpret the literature critically, recognizing the nuanced limitations inherent in RCTs. In this issue of the *Journal*, Kaul and Diamond (4) describe numerous common pitfalls of RCTs. While their examples may stimulate vigorous healthy debate, their report's underlying principles are well reasoned and should be required reading for health care providers and other professionals involved in organizing or interpreting clinical trials.

Many issues addressed by Kaul and Diamond (4) revolve around adequate power. Any RCT should require a prespecified primary end point and a sample size powered to demonstrate either superiority or noninferiority for the new treatment. Small underpowered trials are prone to either alpha error (false positives, with potential publication bias) (5) or beta error (false negatives) (6). Importantly, the magnitude of the treatment effect assumed should be based on realistic expectations. For example, there is little reason to expect that N-acetylcysteine should virtually eliminate radiocontrast nephropathy. Nonetheless, an underpowered trial demonstrating an 89% reduction (7) prompted several additional (underpowered) trials, not surprisingly resulting in conflicting outcomes and publication bias (8). We doubt that Bayesian methods add much insight to standard frequentist methods, except conceptually they do help shrink implausibly large treatment effects in small trials. Bayesian

From the *Columbia University Medical Center and the Cardiovascular Research Foundation, New York, New York; and the †London School of Hygiene and Tropical Medicine, London, United Kingdom. Dr. Stone has served on the advisory board for Boston Scientific and Abbott Vascular, and has received research support from Atrium and Therox. Dr. Pocock has received funding for research, educational, and advisory activities from several pharmaceutical and medical device companies.

Manuscript received June 17, 2009; accepted June 21, 2009.

analysis with hierarchical modeling from serial trials may, however, permit smaller studies to be performed while preserving power, thus enhancing clinical trial efficiency (9), although at the risk of complicating clinical interpretation.

The principal interpretation of an RCT focuses on the single powered primary end point; additional observations should be hypothesis generating unless pre-specified and appropriately powered with statistical adjustment to preserve alpha. Overinterpreting the results of secondary end points or inherently underpowered subgroups risks distortive conclusions, especially regarding rare safety end points, for example, death or stent thrombosis, which can falsely guide patient care decisions or subsequent investigative directions. For example, the unexpected observation of improved survival with the complement inhibitor pexelizumab during primary percutaneous coronary intervention in acute myocardial infarction (AMI) in a pilot trial that failed to demonstrate infarct size reduction (10) led to the costly performance of an 8,500-patient RCT powered for mortality, which was stopped prematurely for futility (11). The interpretation of composite end points may be challenging if the component end points are not uniformly affected by treatment, or are of differing clinical importance. Few end points, however, are judged as equally devastating as death (12), and rarely can trials be powered for mortality (the GUSTO [Global Utilization of Streptokinase and Tissue-Type Plasminogen Activator for Occluded Coronary Arteries] trial, for example, required >40,000 patients) (13). Presentation of the individual component event rates is thus vital, and sensible interpretation depends on the pattern observed. Weighted composite end points are conceptually interesting, but given their subjectivity, are best reserved for

Abbreviations and Acronyms

AMI = acute myocardial infarction

RCT = randomized controlled trial

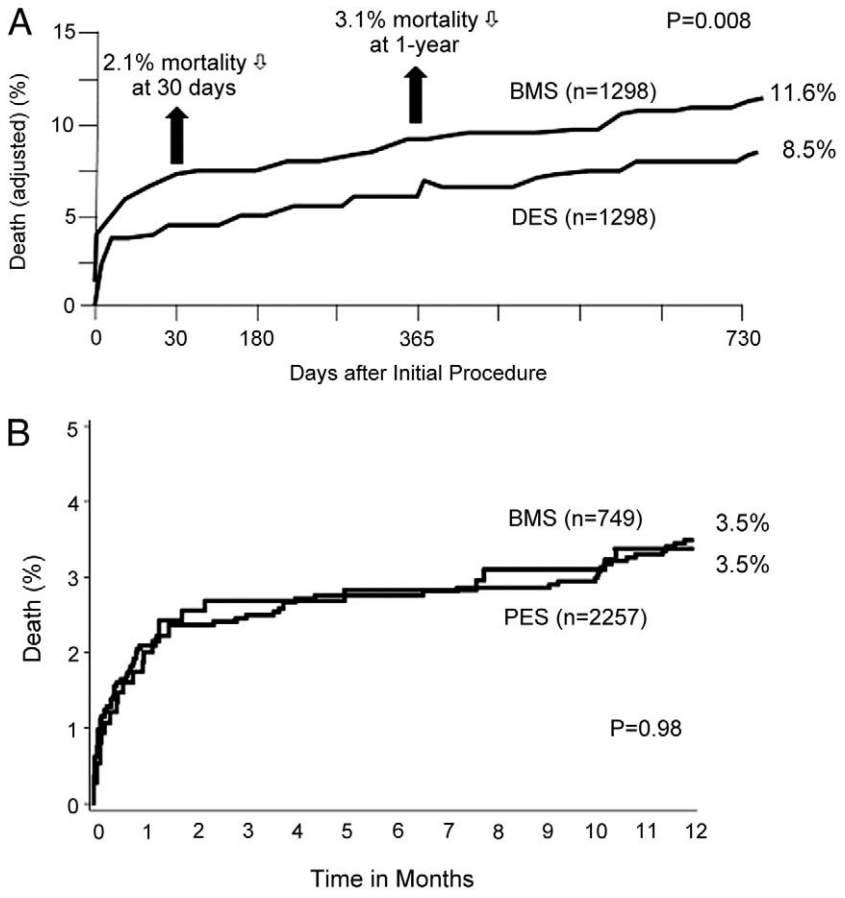

Figure 1 **Comparative Mortality Rates for DES and BMS in Acute Myocardial Infarction**

(A) Two-year mortality among 1,298 propensity-matched pairs of patients with ST-segment elevation myocardial infarction treated at physician discretion with either paclitaxel or sirolimus drug-eluting stents (DES) or bare-metal stents (BMS) at 21 Massachusetts hospitals. Significantly lower mortality was present with DES, mostly by 30 days, a time before any known benefits of DES compared with BMS. This finding may well be due to residual confounding by unmeasured variables in this nonrandomized study. Adapted from Mauri et al. (2). **(B)** One-year mortality among 3,006 patients with ST-segment elevation myocardial infarction randomly allocated in a 3:1 ratio to either paclitaxel-eluting stents (PES) or BMS at 123 hospitals in 11 countries. Mortality rates with DES and BMS were nearly identical at 30 days and 1 year. Adapted from Stone et al. (3).

sensitivity analysis (4). Subgroup analyses are best examined in large trials, with pre-specification of those few groupings of genuine interest, should be explored by statistical interaction paying regard to multiple comparisons, and should only be considered hypothesis generating, requiring prospective validation in future studies.

While we agree with most of Kaul and Diamond's arguments (4), several examples contained internal inconsistencies. For example, the Stent-PAMI (Stent–Primary Angioplasty for Myocardial Infarction) trial was too small ($n = 900$) to raise serious concern regarding increased mortality with stenting compared with balloon angioplasty for AMI (especially with $p > 0.05$). Subsequent trials ($n = 6,922$ total) confirmed near-identical survival rates (14). Similarly, regarding the comparative safety and efficacy of drug-eluting versus bare-metal stents, studies in >12,000 randomized patients and >450,000 registry patients have now demonstrated that mortality and myocardial infarction rates are comparable or reduced with drug-eluting stents (15–17). Kaul and Diamond (4) note that bleeding and myocardial infarction rates tracked discordantly in 2 trials in which bivalirudin was compared with heparin plus glycoprotein IIb/IIIa inhibitors, arguing they should not have been combined in a composite end point. Yet, 1-year mortality with bivalirudin was significantly reduced among patients with AMI (17) and also when pooled across 3 trials covering the spectrum of coronary artery disease in >23,000 patients, reflecting the long-term impact of major hemorrhagic complications on mortality (18). No single study or isolated finding from an RCT should dictate the practice of medicine; even large RCTs must be considered in context with the results from other high-quality studies.

Several additional key aspects of RCTs were not discussed by Kaul and Diamond (4). 1) Multicenter trials afford greater credence and generalizability than single-center studies. 2) Double-blind trials are preferred to single-blind trials, which are preferred to fully unblinded studies. However, unblinded trials, while often logistically unavoidable, can yield valid results if there is high compliance with protocol procedures, appropriate efforts to minimize bias, and use of blinded end point committees and core laboratories. 3) Superiority trials provide stronger evidence than noninferiority trials with active controls; the latter are limited by the acceptability of the chosen noninferiority margin (delta) and creep, which can occur when serial noninferiority trials are performed, resulting in falsely declaring an inferior treatment equivalent to a prior standard (19). 4) Whether the patients enrolled and study processes are generalizable to other settings requires careful consideration. For example, an RCT evaluating an anticoagulant in acute coronary syndrome patients managed conservatively has little relevance to patients managed invasively. 5) The quality of RCTs varies greatly in terms of their design, conduct, and reporting. Studies of lesser quality and smaller scope should not carry equal weight to large-scale comprehensive trials just because they are "randomized." Also, the benefits from the robust quality control often present in industry-sponsored trials, which for regulatory reasons employ more rigorous data collection, monitoring, and independent oversight committees and core laboratories than do most investigator-sponsored studies, must be weighed against concerns of potential bias in trial design, implementation, and reporting.

Also of concern is the ever-increasing number of meta-analyses being published (often in high-quality journals), given their inherent limitations and variable quality. Meta-analyses are frequently constructed from numerous underpowered and flawed trials, leading to questionable "meta-conclusions" that may well be wrong when compared with a subsequent large RCT (20). Random effects models are often utilized to account for heterogeneity between studies that should not have been grouped together in the first place, and provide undue weight to the results from small trials (21). Network meta-analyses incorporate evidence from studies that do not directly compare the treatments of interest (22), representing in many cases a statistical leap too far. Meta-regression is a seriously flawed instrument; such observational associations of variations in effect between trials are inevitably a very weak form of evidence (23). Despite these limitations, meta-analysis and meta-regression are receiving unwarranted attention by journals, guidelines committees, and payors.

We hope that the paper by Kaul and Diamond (4) and this commentary will spur improved design of future RCTs and thoughtful critical appraisal by caregivers, editors, and regulatory bodies. While it is true that "a p value is no substitute for a brain," neither is ignorance of sound clinical investigation and statistical principles. Translating statistical findings into plain English will facilitate the appropriate clinical interpretation of RCTs (24). The implementation of evidence from well-designed scientific studies into the everyday practice of medicine is essential if clinical medicine is to realize its full potential.

Reprint requests and correspondence: Dr. Gregg W. Stone, Columbia University Medical Center, 161 Fort Washington, New York, New York 10023. E-mail: gs2184@columbia.edu.

REFERENCES

1. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? N Engl J Med 2000;342:1907–9.
2. Mauri L, Silbaugh TS, Garg P, et al. Drug-eluting or bare-metal stents for acute myocardial infarction. N Engl J Med 2008;359:1330–42.
3. Stone GW, Lansky AJ, Pocock SJ, et al. Paclitaxel-eluting stents versus bare-metal stents in acute myocardial infarction. N Engl J Med 2009;360:1946–59.
4. Kaul S, Diamond GA. Trial and error: how to avoid commonly encountered limitations of published clinical trials. J Am Coll Cardiol 2010;55:415–27.
5. DeMaria AN. Publication bias and journals as policemen. J Am Coll Cardiol 2004;44:1707–8.
6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.

7. Tepel M, Van der Giet M, Schwarzfeld C, et al. Prevention of radiographic contrast agent-induced reductions in renal function by acetylcysteine. N Engl J Med 2000;343:180–4.
8. Vaitkus PT, Brar C. N-acetylcysteine in the prevention of contrast-induced nephropathy: publication bias perpetuated by meta-analyses. Am Heart J 2007;153:275–80.
9. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Available at: http://www.fda.gov/cdrh/osb/guidance/1601.pdf. Accessed June 12, 2009.
10. Granger CB, Mahaffey KW, Weaver WD, et al. Pexelizumab, an anti-C5 complement antibody, as adjunctive therapy to primary percutaneous coronary intervention in acute myocardial infarction. Circulation 2003;108:1184–90.
11. The APEX AMI Investigators. Pexelizumab for acute ST-elevation myocardial infarction in patients undergoing primary percutaneous coronary intervention. JAMA 2007;297:43–51.
12. Tengs TO, Lin TH. A meta-analysis of quality-of-life estimates for stroke. Pharmacoeconomics 2003;21:191–200.
13. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993;329:673–82.
14. De Luca G, Suryapranata H, Stone GW, et al. Coronary stenting versus balloon angioplasty for acute myocardial infarction. Int J Cardiol 2007;119:306–9.
15. Kirtane AJ, Gupta A, Iyengar S, et al. Safety and efficacy of drug-eluting and bare metal stents: comprehensive meta-analysis of randomized trials and observational studies. Circulation 2009;119:3198–206.
16. Douglas PS, Brennan JM, Anstrom KJ, et al. Clinical effectiveness of coronary stents in elderly persons. J Am Coll Cardiol 2009;53:1629–41.
17. Stone GW, Witzenbichler B, Guagliumi G, et al. Bivalirudin during primary PCI in acute myocardial infarction. N Engl J Med 2008;358:2218–30.
18. Doyle BJ, Rihal CS, Gastineau DA, Holmes DR. Bleeding, blood transfusion, and increased mortality after percutaneous coronary intervention. J Am Coll Cardiol 2009;53:2019–27.
19. Ware JH, Antman EM. Equivalence trials. N Engl J Med 1997;337:1159–61.
20. LeLorier J, Grégoire G, Benhaddad A, et al. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536–42.
21. Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet 1992;338:1127–30.
22. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313–24.
23. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23:1663–82.
24. Pocock SJ, Ware JH. Translating statistical findings into plain English. Lancet 2009;373:1926–8.

Key Words: results ■ megatrials ■ interpret.